Matemáticas para la Universidad

TEMA 6

Estadística Descriptiva

La estadística descriptiva nos permite **resumir**, **organizar** y **visualizar** datos.

Es la base de toda investigación científica, social o empresarial: sin describir bien los datos, no podemos interpretarlos ni tomar decisiones.

Curso de Nivelación Universitaria

11 de junio de 2026

Índice

1. Conceptos básicos: población, muestra y variable **2**
1.1. Tipos de variables . . . 2

2. El símbolo sumatorio Σ **3**
2.1. Definición y notación . . . 3
2.2. Propiedades del sumatorio . . . 3
2.3. Aplicación a la media y la varianza . . . 4
2.4. Sumatorios dobles . . . 5
2.5. Sumatorios especiales . . . 5

3. Organización de datos: tablas de frecuencias **7**
3.1. Datos agrupados en intervalos . . . 7

4. Representaciones gráficas **8**
4.1. Diagrama de barras e histograma . . . 9
4.2. Polígono de frecuencias y ojiva . . . 9
4.3. Diagrama de sectores (tarta) . . . 9

5. Medidas de posición central **10**
5.1. Media aritmética . . . 10
5.2. Mediana . . . 11
5.3. Moda . . . 12

6. Medidas de posición no central: cuantiles **13**
6.1. Diagrama de caja y bigotes (box plot) . . . 13

7. Medidas de dispersión **14**
7.1. Rango o recorrido . . . 14
7.2. Varianza y desviación típica . . . 14
7.3. Coeficiente de variación . . . 16

8. Medidas de forma: asimetría y curtosis **16**
8.1. Asimetría . . . 16
8.2. Curtosis . . . 17

9. Tipificación de variables **17**

10. Distribuciones bidimensionales: correlación y regresión **18**
10.1. Nube de puntos y correlación . . . 19
10.2. Recta de regresión . . . 19

11. Problemas de aplicación **21**

12. Soluciones de las actividades seleccionadas **22**

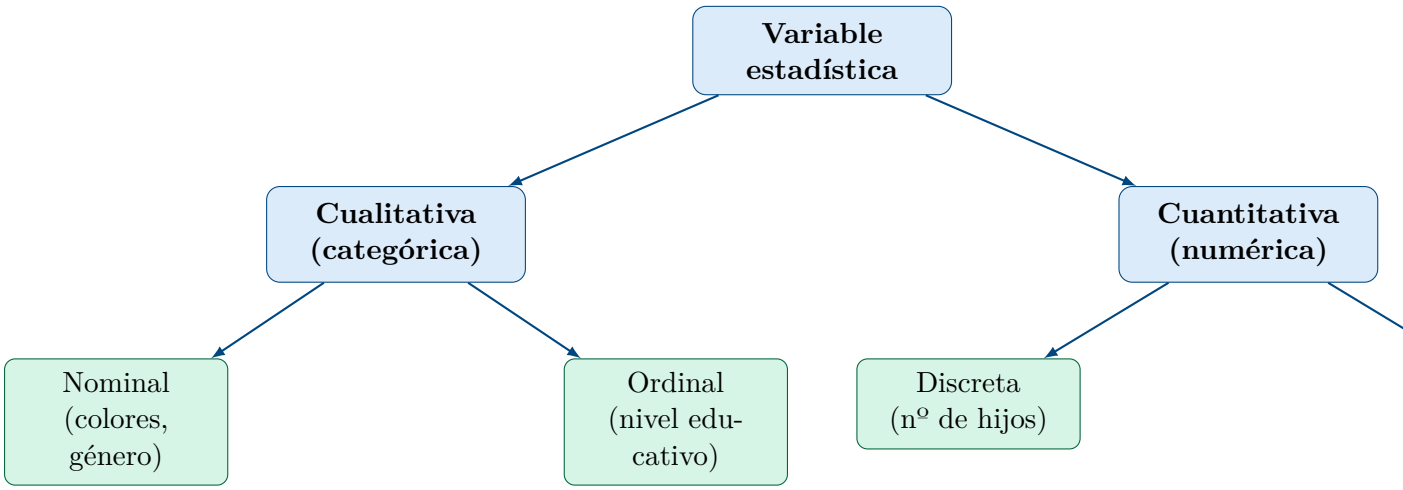
13. Resumen del tema **23**

1. Conceptos básicos: población, muestra y variable

Conceptos fundamentales

- **Población**: conjunto completo de individuos u objetos sobre los que se desea obtener información.
- **Muestra**: subconjunto representativo de la población que se estudia en la práctica.
- **Individuo** (o unidad estadística): cada elemento de la población o muestra.
- **Variable estadística**: característica que se mide o registra en cada individuo.
- **Dato**: valor concreto de una variable para un individuo.

1.1. Tipos de variables



Actividad 1.1 — Tipos de variables

Clasifica las siguientes variables (cualitativa nominal / cualitativa ordinal / cuantitativa discreta / cuantitativa continua):

1. Número de hermanos
2. Temperatura corporal
3. Color de ojos
4. Nota de un examen (suspenso / aprobado / notable / sobresaliente)
5. Número de accidentes en una autopista al mes
6. Marca de un móvil
7. Tiempo empleado en correr 100 metros
8. Satisfacción del cliente (1 a 5 estrellas)

2. El símbolo sumatorio Σ

A partir de ahora usaremos con frecuencia el **símbolo sumatorio** Σ (letra griega mayúscula sigma). Merece la pena dedicarle un apartado propio porque aparece en casi todas las fórmulas estadísticas.

2.1. Definición y notación

Símbolo sumatorio

Dado un conjunto de valores $a_1, a_2, \ldots, a_n$, se define:

$$\sum_{i=1}^{n} a_i = a_1 + a_2 + a_3 + \cdots + a_n$$

- i es el **índice** de sumación (variable muda: el resultado no depende de la letra elegida).
- $i = 1$ es el **límite inferior** (valor inicial del índice).
- n es el **límite superior** (valor final del índice).
- a_i es el **término general** de la suma.

Lectura del sumatorio

$$\sum_{i=1}^{4} i^2 = 1^2 + 2^2 + 3^2 + 4^2 = 1 + 4 + 9 + 16 = 30$$

$$\sum_{i=1}^{5} (2i - 1) = 1 + 3 + 5 + 7 + 9 = 25 \quad \text{(suma de los 5 primeros impares)}$$

$$\sum_{i=0}^{3} \frac{1}{2^i} = \frac{1}{1} + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} = \frac{15}{8}$$

El índice es una variable muda

El nombre del índice no importa; todas estas expresiones son iguales:

$$\sum_{i=1}^{n} a_i = \sum_{j=1}^{n} a_j = \sum_{k=1}^{n} a_k$$

En estadística se usa habitualmente i para recorrer individuos.

2.2. Propiedades del sumatorio

Las siguientes propiedades permiten simplificar y manipular expresiones con sumatorios. Son la base de todas las demostraciones estadísticas.

Propiedades fundamentales del sumatorio

Sean a_i, b_i valores cualesquiera, c una constante y $n \in \mathbb{N}$:

P1. Suma de sumas:

$$\sum_{i=1}^{n}(a_i + b_i) = \sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i$$

P2. Factor constante:

$$\sum_{i=1}^{n} c\, a_i = c \sum_{i=1}^{n} a_i$$

P3. Suma de una constante:

$$\sum_{i=1}^{n} c = n \cdot c$$

P4. Combinación lineal (P1 + P2 juntas):

$$\sum_{i=1}^{n}(\alpha\, a_i + \beta\, b_i) = \alpha \sum_{i=1}^{n} a_i + \beta \sum_{i=1}^{n} b_i$$

P5. Descomposición del rango:

$$\sum_{i=1}^{n} a_i = \sum_{i=1}^{m} a_i + \sum_{i=m+1}^{n} a_i \quad (1 \le m < n)$$

Lo que NO se puede hacer

Estos errores son muy frecuentes — ¡evítalos!

$$\sum_{i=1}^{n} a_i \cdot b_i \neq \left(\sum_{i=1}^{n} a_i\right)\cdot\left(\sum_{i=1}^{n} b_i\right) \qquad \sum_{i=1}^{n} a_i^2 \neq \left(\sum_{i=1}^{n} a_i\right)^2$$

El sumatorio **no** se distribuye sobre el producto ni sobre la potencia.

Contraejemplo rápido con $a = (1, 2)$, $b = (3, 4)$*:*
$\sum a_i b_i = 1 \cdot 3 + 2 \cdot 4 = 11$ pero $\left(\sum a_i\right)\left(\sum b_i\right) = 3 \cdot 7 = 21 \neq 11$.

2.3. Aplicación a la media y la varianza

La suma de las desviaciones a la media es siempre cero

$$\sum_{i=1}^{n}(x_i - \bar{x}) \overset{\text{P1}}{=} \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} \overset{\text{P2,P3}}{=} n\bar{x} - n\bar{x} = 0$$

La media es el *punto de equilibrio* de los datos: las desviaciones positivas y negativas se cancelan exactamente.

Fórmula alternativa de la varianza

Demostración de que $s^2 = \overline{x^2} - \bar{x}^2$:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)$$

$$\overset{\text{P1,P2,P3}}{=} \frac{1}{n}\sum x_i^2 - \frac{2\bar{x}}{n}\sum x_i + \bar{x}^2$$

$$= \overline{x^2} - 2\bar{x}\cdot\bar{x} + \bar{x}^2 = \overline{x^2} - \bar{x}^2 \qquad \checkmark$$

2.4. Sumatorios dobles

En estadística bidimensional (correlación, regresión) aparecen sumatorios con dos índices:

Sumatorio doble

$$\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij} = \sum_{i=1}^{m}\left(\sum_{j=1}^{n} a_{ij}\right)$$

Se calcula desde *dentro hacia afuera*: primero el sumatorio en j (con i fijo) y luego el de i.

Ejemplo

$$\sum_{i=1}^{2}\sum_{j=1}^{3} ij = \sum_{i=1}^{2}(i\cdot 1 + i\cdot 2 + i\cdot 3) = \sum_{i=1}^{2} 6i = 6\cdot 1 + 6\cdot 2 = 18$$

2.5. Sumatorios especiales

Resultados cerrados que evitan sumar término a término:

Fórmula	**Ejemplo con $n = 5$**
$\sum_{i=1}^{n} 1 = n$	$1{+}1{+}1{+}1{+}1 = 5$
$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$	$1{+}2{+}3{+}4{+}5 = 15$
$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$	$1{+}4{+}9{+}16{+}25 = 55$
$\sum_{i=1}^{n} i^3 = \left[\frac{n(n+1)}{2}\right]^2$	$1{+}8{+}27{+}64{+}125 = 225 = 15^2$
$\sum_{i=0}^{n} r^i = \frac{1-r^{n+1}}{1-r}$ $(r \neq 1)$	$\sum_{i=0}^{4} 2^i = 31$

Actividad 2.1 — El símbolo sumatorio

1. Desarrolla y calcula:

   a) $\sum_{i=1}^{5}(3i+1)$

   b) $\sum_{i=1}^{4} i(i-1)$

   c) $\sum_{i=2}^{6} \frac{i}{i+1}$

   d) $\sum_{i=0}^{4}(-1)^i$

2. Usa las fórmulas cerradas para calcular sin desarrollar:

   a) $1+2+3+\cdots+100$

   b) $1^2+2^2+3^2+\cdots+10^2$

   c) $1+3+9+27+81+243$ *(suma geométrica de razón 3)*

3. Escribe en notación de sumatorio:

   a) $x_1f_1+x_2f_2+\cdots+x_kf_k$

   b) $(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2$

   c) $\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16}$

4. Indica si son verdaderas (V) o falsas (F). Si son falsas, da un contraejemplo con $n=2$, $a_1=1$, $a_2=2$:

   a) $\sum_{i=1}^{n}(a_i+3)=3n+\sum_{i=1}^{n}a_i$

   b) $\sum_{i=1}^{n}a_i^2=\left(\sum_{i=1}^{n}a_i\right)^2$

   c) $\sum_{i=1}^{n}\frac{a_i}{b_i}=\frac{\sum_{i=1}^{n}a_i}{\sum_{i=1}^{n}b_i}$

   d) $\sum_{i=1}^{n}c\,a_i=c\sum_{i=1}^{n}a_i$

5. Demuestra con las propiedades del sumatorio que, si $y_i=ax_i+b$ (transformación lineal de los datos):

   $$\bar{y}=a\bar{x}+b$$

   Pista: expresa $\bar{y}$ como $\frac{1}{n}\sum y_i$ y aplica P1, P2 y P3.

3. Organización de datos: tablas de frecuencias

Tipos de frecuencia

Para un conjunto de n datos donde el valor x_i aparece f_i veces:

- f_i: **frecuencia absoluta** de x_i (número de veces que aparece).
- $F_i = \sum_{j=1}^{i} f_j$: **frecuencia absoluta acumulada**.
- $h_i = \dfrac{f_i}{n}$: **frecuencia relativa** (tanto por uno).
- $H_i = \sum_{j=1}^{i} h_j$: **frecuencia relativa acumulada**.

¡Atención!

Siempre se cumple: $\sum_i f_i = n$ y $\sum_i h_i = 1$.

Tabla de frecuencias — datos discretos

Las notas de 20 alumnos en un test (de 0 a 10) son:

5, 7, 6, 8, 5, 9, 6, 7, 5, 6, 8, 7, 6, 5, 10, 7, 8, 6, 7, 9

x_i (nota)	f_i	F_i	h_i	H_i
5	4	4	0,20	0,20
6	5	9	0,25	0,45
7	5	14	0,25	0,70
8	3	17	0,15	0,85
9	2	19	0,10	0,95
10	1	20	0,05	1,00
Total	**20**		**1,00**	

3.1. Datos agrupados en intervalos

Cuando los datos son continuos o hay muchos valores distintos, se agrupan en **intervalos de clase** o **clases**.

Datos en intervalos

- **Amplitud** de la clase: $a = L_{\text{sup}} - L_{\text{inf}}$
- **Marca de clase**: $c_i = \dfrac{L_{\text{inf}} + L_{\text{sup}}}{2}$ (punto representativo del intervalo)
- Se usan habitualmente intervalos del tipo $[L_i, L_{i+1})$ (cerrado por la izquierda, abierto por la derecha).

Tabla de frecuencias — datos agrupados

Alturas (en cm) de 30 estudiantes universitarios:

155, 162, 168, 171, 174, 158, 163, 169, 172, 175, 160, 165, 170, 173, 176, 161, 166, 170, 173, 177, 162, 167, 1

Intervalo	Marca c_i	f_i	F_i	h_i	H_i
$[155, 160)$	157,5	2	2	0,067	0,067
$[160, 165)$	162,5	7	9	0,233	0,300
$[165, 170)$	167,5	7	16	0,233	0,533
$[170, 175)$	172,5	9	25	0,300	0,833
$[175, 180)$	177,5	4	29	0,133	0,967
$[180, 185)$	182,5	1	30	0,033	1,000
Total		**30**		**1,000**	

Actividad 2.1 — Tabla de frecuencias

Los tiempos (en minutos) que tarda un grupo de 25 personas en llegar al trabajo son:

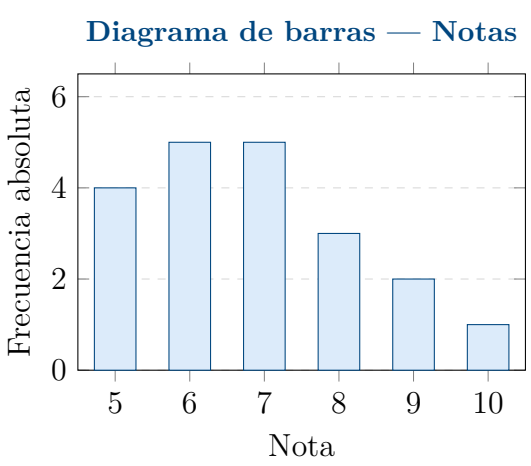
12, 18, 25, 30, 22, 15, 20, 28, 35, 18, 24, 32, 19, 27, 14, 23, 29, 16, 21, 33, 17, 26, 31, 20, 24

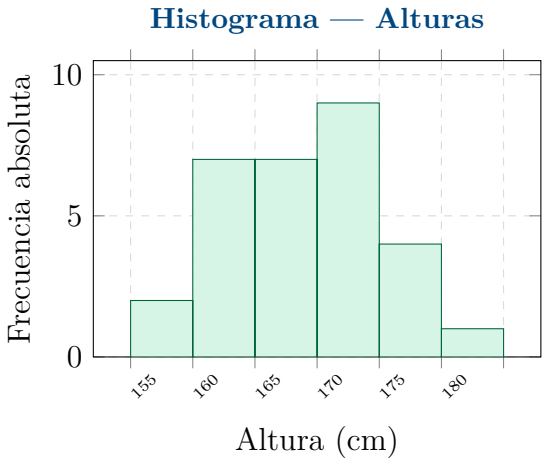
1. Construye una tabla de frecuencias agrupando en intervalos de amplitud 5 (comenzando en 10).
2. Calcula las cuatro frecuencias (f_i, F_i, h_i, H_i).
3. ¿Qué porcentaje de personas tarda entre 20 y 30 minutos?
4. ¿Cuántas personas tardan menos de 25 minutos?

4. Representaciones gráficas

4.1. Diagrama de barras e histograma

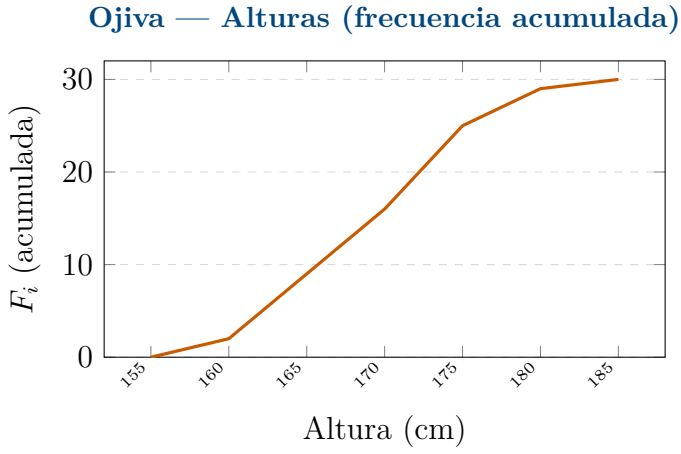
- **Diagrama de barras**: datos discretos o cualitativos. Las barras **no** se tocan.
- **Histograma**: datos continuos agrupados en intervalos. Las barras **sí** se tocan (los intervalos son contiguos). El área de cada barra es proporcional a la frecuencia.





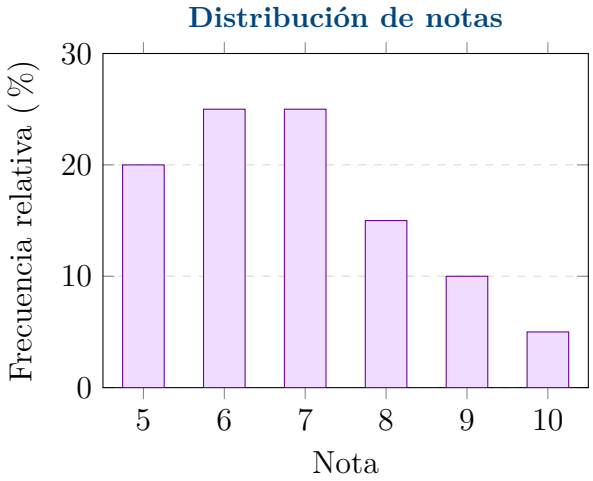
4.2. Polígono de frecuencias y ojiva

- **Polígono de frecuencias**: se unen con segmentos los puntos (c_i, f_i) de las marcas de clase.
- **Ojiva** (o polígono de frecuencias acumuladas): se unen $(L_{\text{sup},i}, F_i)$. Permite leer percentiles directamente.



4.3. Diagrama de sectores (tarta)

Se usa para variables cualitativas o discretas con pocos valores. Cada sector tiene un ángulo proporcional a su frecuencia relativa: $\alpha_i = h_i \times 360ř$.



Actividad 3.1 — Representaciones gráficas

Con los datos de la Actividad 2.1 (tiempos de desplazamiento):

1. Construye el histograma de frecuencias absolutas.
2. Construye la ojiva (frecuencias acumuladas).
3. ¿Qué tiempo aproximado corresponde al percentil 50 (mediana gráfica) según la ojiva?

5. Medidas de posición central

Las medidas de posición central resumen un conjunto de datos con un único valor representativo. Las tres más importantes son la **media**, la **mediana** y la **moda**.

5.1. Media aritmética

Media aritmética

La **media aritmética** $\bar{x}$ de n datos $x_1, x_2, \ldots, x_n$:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Para datos agrupados en tabla de frecuencias:

$$\bar{x} = \frac{\sum_i x_i \cdot f_i}{n} = \sum_i x_i \cdot h_i$$

Para datos en intervalos se usan las marcas de clase c_i:

$$\bar{x} \approx \frac{\sum_i c_i \cdot f_i}{n}$$

Media de las notas

Usando la tabla de frecuencias anterior:

$$\bar{x} = \frac{5\cdot 4 + 6\cdot 5 + 7\cdot 5 + 8\cdot 3 + 9\cdot 2 + 10\cdot 1}{20} = \frac{20+30+35+24+18+10}{20} = \frac{137}{20} = 6{,}85$$

Sensibilidad de la media a los valores extremos

La media se ve muy afectada por los **valores atípicos** (outliers). Si un alumno saca 0 en el ejemplo anterior, la media bajaría considerablemente aunque el resto no haya cambiado.

5.2. Mediana

Mediana

La **mediana** Me es el valor que divide la distribución en dos mitades iguales: el $50,\%$ de los datos queda por debajo y el $50,\%$ por encima.
Para datos sin agrupar (ordenados de menor a mayor):

- Si n es impar: la mediana es el dato central (posición $\frac{n+1}{2}$).
- Si n es par: la mediana es la media de los dos datos centrales (posiciones $\frac{n}{2}$ y $\frac{n}{2}+1$).

Para datos agrupados en intervalos, interpolando:

$$Me = L_i + \frac{\frac{n}{2} - F_{i-1}}{f_i} \cdot a_i$$

donde $[L_i, L_i + a_i)$ es el intervalo que contiene la mediana.

Mediana de datos sin agrupar

Datos (ya ordenados): $5, 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 8, 8, 8, 9, 9, 10$.
$n = 20$ (par) $\Rightarrow$ posiciones 10 y 11. Datos 10.º y 11.º: **7** y **7**.

$$Me = \frac{7+7}{2} = 7$$

5.3. Moda

Moda

La **moda** Mo es el valor (o intervalo) que aparece con mayor frecuencia. Puede haber más de una moda (distribución bimodal, multimodal) o ninguna (todos con la misma frecuencia).
Para datos agrupados en intervalos, la **clase modal** es el intervalo de mayor frecuencia.

Ejemplo

En la tabla de notas: los valores 6 y 7 aparecen ambos 5 veces. La distribución es **bimodal**: $Mo = 6$ y $Mo = 7$.

Actividad 4.1 — Medidas de posición central

1. Calcula la media, mediana y moda del siguiente conjunto de datos:
   3, 7, 5, 9, 4, 7, 8, 7, 2, 6, 5, 9, 7, 3

2. Con los datos agrupados de las alturas (tabla del ejemplo), calcula la media usando marcas de clase.

3. Los salarios mensuales (en euros) de 7 empleados son:
   $1200, 1350, 1400, 1450, 1500, 1550, 8000$.
   Calcula la media y la mediana. ¿Cuál describe mejor el salario típico? ¿Por qué?

4. La mediana de los siguientes datos es 6. Halla x:
   2, 4, 5, x, 8, 9, 11 (datos ordenados)

5. Calcula la media ponderada de las notas de un alumno sabiendo que Matemáticas (nota 7) tiene peso 3, Física (nota 6) tiene peso 2 y Química (nota 8) tiene peso 2.

6. Medidas de posición no central: cuantiles

Cuantiles

Los **cuantiles** son valores que dividen la distribución en partes iguales:

- **Cuartiles** Q_1, Q_2, Q_3: dividen en cuatro partes. $Q_2 = Me$ (la mediana).
- **Deciles** $D_1, D_2, \ldots, D_9$: dividen en diez partes.
- **Percentiles** $P_1, P_2, \ldots, P_{99}$: dividen en cien partes.

Interpretación de los cuartiles

- Q_1 (percentil 25): el 25 % de los datos es menor que Q_1.
- Q_2 (percentil 50): la mediana; el 50, % es menor que Q_2.
- Q_3 (percentil 75): el 75 % de los datos es menor que Q_3.

El **rango intercuartílico** (RIC o IQR) es:

$$\text{RIC} = Q_3 - Q_1$$

y mide la dispersión del 50, % central de los datos.

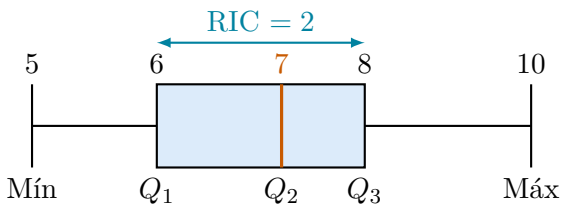
Cálculo de cuartiles

Datos ordenados (20 notas): $5, 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 8, 8, 8, 9, 9, 10$.

- Q_1: mediana de la mitad inferior (primeros 10 datos) = media de datos 5.º y 6.º $= \frac{6+6}{2} = 6$
- $Q_2 = Me = 7$
- Q_3: mediana de la mitad superior (últimos 10 datos) = media de datos 15.º y 16.º $= \frac{8+8}{2} = 8$
- $\text{RIC} = 8 - 6 = 2$

6.1. Diagrama de caja y bigotes (box plot)

El **diagrama de caja y bigotes** resume visualmente la distribución mediante 5 valores: mínimo, Q_1, Q_2, Q_3 y máximo.



Actividad 5.1 — Cuantiles y caja-bigotes

1. Calcula Q_1, Q_2, Q_3 y el RIC para los datos:
   $4, 8, 15, 16, 23, 42, 4, 8, 15, 16$

2. Dibuja el diagrama de caja y bigotes del apartado anterior.

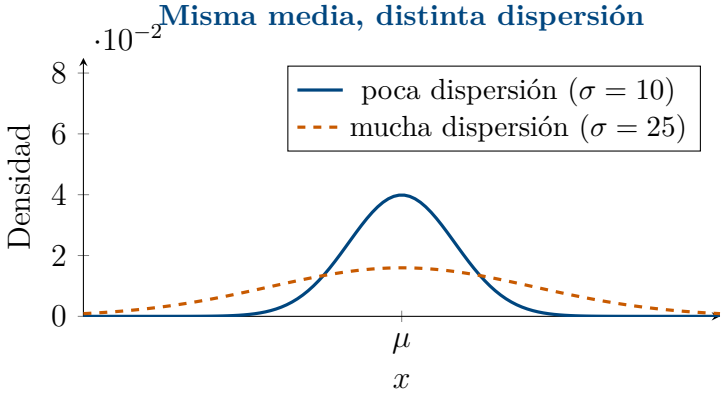
3. Las puntuaciones de dos grupos en un examen son:
   Grupo A: $55, 60, 65, 70, 75, 80, 85$
   Grupo B: $40, 55, 70, 70, 70, 85, 100$
   Calcula Q_1, Q_2, Q_3 y RIC de cada grupo. ¿Qué grupo es más homogéneo?

4. Usando los datos de alturas del ejemplo, interpreta: «Un estudiante está en el percentil 83». ¿Qué significa?

7. Medidas de dispersión

Las medidas de posición central no son suficientes para describir una distribución: dos conjuntos pueden tener la misma media y ser muy distintos en cuanto a su **variabilidad**.

Misma media, distinta dispersión



7.1. Rango o recorrido

$$R = x_{\text{máx}} - x_{\text{mín}}$$

Simple pero muy sensible a valores extremos.

7.2. Varianza y desviación típica

Varianza y desviación típica

La **varianza** mide el promedio de las desviaciones al cuadrado respecto a la media:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2$$

Para datos con frecuencias:

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2 \cdot f_i}{n} = \frac{\sum_i x_i^2 \cdot f_i}{n} - \bar{x}^2$$

La **desviación típica** (o desviación estándar) es:

$$s = \sqrt{s^2}$$

Tiene las mismas unidades que los datos, lo que la hace más interpretable que la varianza.

Varianza poblacional vs. muestral

En estadística inferencial (cuando la muestra no es toda la población), se usa $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$ (divisor $n-1$) para obtener un estimador insesgado. En estadística descriptiva se usa n.

Varianza y desviación típica de las notas

$\bar{x} = 6{,}85$. Tabla auxiliar de cuadrados de desviaciones:

x_i	$(x_i - \bar{x})^2$	f_i	$(x_i - \bar{x})^2 \cdot f_i$
5	$(-1{,}85)^2 = 3{,}4225$	4	13,690
6	$(-0{,}85)^2 = 0{,}7225$	5	3,6125
7	$(+0{,}15)^2 = 0{,}0225$	5	0,1125
8	$(+1{,}15)^2 = 1{,}3225$	3	3,9675
9	$(+2{,}15)^2 = 4{,}6225$	2	9,245
10	$(+3{,}15)^2 = 9{,}9225$	1	9,9225
		20	**40,55**

$$s^2 = \frac{40{,}55}{20} = 2{,}0275 \qquad s = \sqrt{2{,}0275} \approx 1{,}42$$

Fórmula alternativa $s^2 = \overline{x^2} - \bar{x}^2$:

$$\overline{x^2} = \frac{25 \cdot 4 + 36 \cdot 5 + 49 \cdot 5 + 64 \cdot 3 + 81 \cdot 2 + 100 \cdot 1}{20} = \frac{979}{20} = 48{,}95$$

$$s^2 = 48{,}95 - 6{,}85^2 = 48{,}95 - 46{,}9225 = 2{,}0275 \checkmark$$

7.3. Coeficiente de variación

Coeficiente de variación

El **coeficiente de variación** (CV) expresa la desviación típica como porcentaje de la media. Permite comparar la dispersión de distribuciones con distintas unidades o medias:

$$CV = \frac{s}{\bar{x}} \times 100\,\%$$

Ejemplo

Dos grupos:

- Grupo A: $\bar{x} = 50$, $s = 5 \Rightarrow CV = 10\,\%$
- Grupo B: $\bar{x} = 200$, $s = 15 \Rightarrow CV = 7{,}5\,\%$

Aunque $s_B > s_A$, el grupo B es **relativamente más homogéneo**.

Actividad 6.1 — Medidas de dispersión

1. Calcula la varianza y la desviación típica para: 2, 4, 4, 4, 5, 5, 7, 9. Usa la fórmula alternativa $s^2 = \overline{x^2} - \bar{x}^2$.
2. Con los datos de las notas del ejemplo, comprueba los resultados usando directamente la definición.
3. Los pesos (en kg) de dos equipos de baloncesto son:
   Equipo A: $78, 82, 85, 80, 79, 83, 81$
   Equipo B: $70, 90, 76, 95, 72, 88, 84$
   Calcula el CV de cada equipo y compara la homogeneidad.
4. Una variable tiene $\bar{x} = 20$ y $s^2 = 16$. Si a todos los datos se les suma 5, ¿cambia la varianza? Si se multiplican por 2, ¿cómo cambia la desviación típica? Razona.

8. Medidas de forma: asimetría y curtosis

8.1. Asimetría

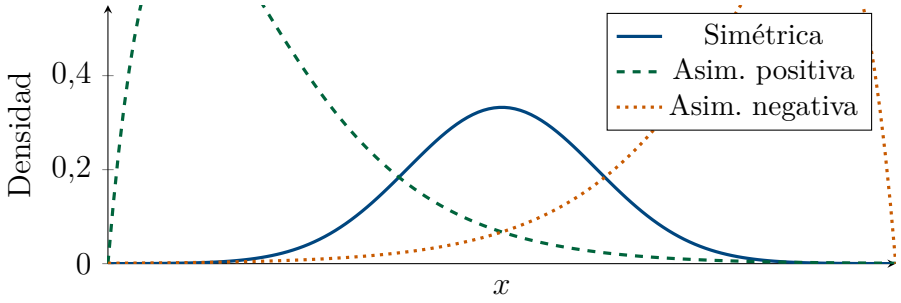
Coeficiente de asimetría de Pearson

$$g_1 = \frac{\bar{x} - Mo}{s} \qquad \text{(aproximado)}$$

O el coeficiente estandarizado (momento de orden 3):

$$\gamma_1 = \frac{\sum_i (x_i - \bar{x})^3 \cdot f_i}{n \cdot s^3}$$

- $\gamma_1 = 0$: distribución **simétrica**
- $\gamma_1 > 0$: **asimetría positiva** (cola larga a la derecha)
- $\gamma_1 < 0$: **asimetría negativa** (cola larga a la izquierda)



8.2. Curtosis

La **curtosis** mide el apuntamiento de la distribución respecto a la normal:

$$\gamma_2 = \frac{\sum_i (x_i - \bar{x})^4 \cdot f_i}{n \cdot s^4} - 3$$

- $\gamma_2 = 0$: distribución **mesocúrtica** (como la normal)
- $\gamma_2 > 0$: **leptocúrtica** (más apuntada, colas pesadas)
- $\gamma_2 < 0$: **platicúrtica** (más aplanada, colas ligeras)

Actividad 7.1 — Forma de la distribución

1. Para la distribución de notas, el coeficiente de Pearson da $g_1 = \frac{6{,}85-7}{1{,}42} \approx -0{,}106$. ¿Es simétrica, con asimetría positiva o negativa? Interpreta.
2. Los ingresos de un país tienen asimetría fuertemente positiva. ¿Qué implica eso sobre la relación entre media y mediana?
3. Dibuja un boceto de una distribución con asimetría negativa y señala aproximadamente dónde estarían la media, la mediana y la moda.

9. Tipificación de variables

Variable tipificada

Dada una variable X con media $\bar{x}$ y desviación típica s, la **puntuación típica** (o *z-score*) de un dato x_i es:

$$z_i = \frac{x_i - \bar{x}}{s}$$

La variable tipificada Z tiene media 0 y desviación típica 1.

¿Para qué sirve la tipificación?

- **Comparar individuos de distintas distribuciones**: ¿Quién está relativamente mejor, alguien con nota 7/10 en un examen fácil o alguien con nota 6/10 en uno difícil?
- **Detectar outliers**: convencionalmente, un dato con $|z| > 3$ se considera atípico.
- **Conectar con la distribución normal estándar** (Tema 5).

Ejemplo

María saca 75 en Matemáticas ($\bar{x} = 65$, $s = 10$) y 80 en Física ($\bar{x} = 78$, $s = 6$).

$$z_{\text{Mates}} = \frac{75 - 65}{10} = 1{,}0 \qquad z_{\text{Física}} = \frac{80 - 78}{6} \approx 0{,}33$$

María está **relativamente mejor en Matemáticas** (1 desviación típica por encima de la media) que en Física.

Actividad 8.1 — Tipificación

1. En un examen con $\bar{x} = 58$ y $s = 12$, tipifica las puntuaciones: 70, 46, 58, 82.
2. Un alumno obtiene $z = -1{,}5$ en Química. Si $\bar{x} = 65$ y $s = 10$, ¿qué nota obtuvo?
3. Dos estudiantes compiten por una beca. El criterio es la puntuación tipificada conjunta de dos pruebas:
   Ana: Prueba 1 $= 85$ ($\bar{x} = 80, s = 8$); Prueba 2 $= 72$ ($\bar{x} = 68, s = 10$)
   Luis: Prueba 1 $= 90$ ($\bar{x} = 80, s = 8$); Prueba 2 $= 68$ ($\bar{x} = 68, s = 10$)
   ¿Quién tiene mejor puntuación tipificada media?

10. Distribuciones bidimensionales: correlación y regresión

Hasta ahora hemos estudiado **una** variable. En este apartado estudiamos la relación entre **dos** variables.

10.1. Nube de puntos y correlación

Covarianza y coeficiente de correlación

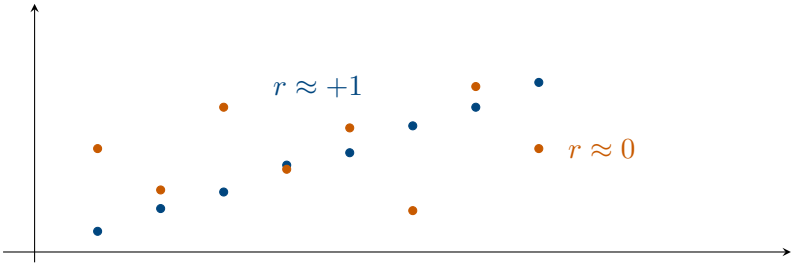
Para n pares de datos (x_i, y_i):

$$\text{Cov}(X,Y) = \overline{xy} - \bar{x}\cdot\bar{y} = \frac{\sum_i x_i y_i \cdot f_i}{n} - \bar{x}\cdot\bar{y}$$

El **coeficiente de correlación de Pearson**:

$$r = \frac{\text{Cov}(X,Y)}{s_X \cdot s_Y} \qquad -1 \le r \le 1$$

Interpretación de r



Interpretación de r

- $r = +1$: correlación lineal positiva perfecta
- $0{,}7 \le r < 1$: correlación positiva fuerte
- $0{,}3 \le r < 0{,}7$: correlación positiva moderada
- $|r| < 0{,}3$: correlación débil o nula
- $r = -1$: correlación lineal negativa perfecta

10.2. Recta de regresión

Recta de regresión de Y sobre X

La recta que minimiza la suma de los cuadrados de los residuos verticales (criterio de mínimos cuadrados):

$$y = a + bx \qquad \text{donde} \qquad b = \frac{\text{Cov}(X,Y)}{s_X^2}, \quad a = \bar{y} - b\bar{x}$$

Regresión: horas de estudio vs. nota

Horas (x)	2	3	4	5	6	7
Nota (y)	4	5	6	7	7	9

$$\bar{x} = \frac{2+3+4+5+6+7}{6} = \frac{27}{6} = 4{,}5 \qquad \bar{y} = \frac{4+5+6+7+7+9}{6} = \frac{38}{6} \approx 6{,}33$$

$$\overline{xy} = \frac{2\cdot 4 + 3\cdot 5 + 4\cdot 6 + 5\cdot 7 + 6\cdot 7 + 7\cdot 9}{6} = \frac{8+15+24+35+42+63}{6} = \frac{187}{6} \approx 31{,}17$$

$$\mathrm{Cov}(X,Y) = 31{,}17 - 4{,}5 \cdot 6{,}33 = 31{,}17 - 28{,}48 = 2{,}69$$

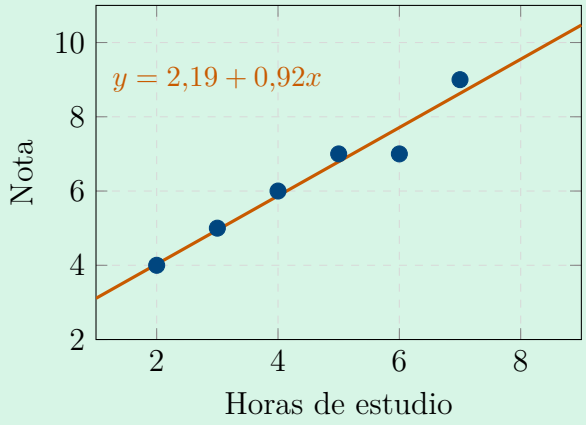
$$s_X^2 = \overline{x^2} - \bar{x}^2 = \frac{4+9+16+25+36+49}{6} - 4{,}5^2 = 23{,}17 - 20{,}25 = 2{,}92$$

$$b = \frac{2{,}69}{2{,}92} \approx 0{,}92 \qquad a = 6{,}33 - 0{,}92 \cdot 4{,}5 = 6{,}33 - 4{,}14 = 2{,}19$$

$$\boxed{y = 2{,}19 + 0{,}92\,x}$$

Predicción para 8 horas: $y = 2{,}19 + 0{,}92 \cdot 8 = 9{,}55$

Recta de regresión



Actividad 9.1 — Correlación y regresión

1. Los datos de temperatura (°C) y ventas de helados (unidades) en 6 días son:

Temperatura (x)	20	24	28	30	35	38
Ventas (y)	80	110	150	170	210	250

   a) Calcula $\bar{x}$, $\bar{y}$, s_X, s_Y y $\mathrm{Cov}(X,Y)$.

   b) Halla el coeficiente de correlación r e interprétalo.

   c) Encuentra la recta de regresión $y = a + bx$.

   d) Predice las ventas para una temperatura de 32°C.

2. ¿Puede concluirse que la temperatura **causa** un aumento en las ventas de helados? ¿Por qué? (Reflexiona sobre la diferencia entre correlación y causalidad.)

11. Problemas de aplicación

Problema 1 — Análisis completo de datos

Las calificaciones de 30 alumnos en Matemáticas son:

$$4, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 9, 9, 9, 9, 10, 10, 3, 5, 6, 7, 8$$

1. Construye la tabla de frecuencias completa.
2. Calcula la media, mediana y moda.
3. Calcula la varianza, la desviación típica y el CV.
4. Calcula los cuartiles y el RIC.
5. Dibuja el histograma y el diagrama de caja y bigotes.
6. Comenta la forma de la distribución.

Problema 2 — Comparación de grupos

Dos métodos de enseñanza se prueban con 10 alumnos cada uno. Las notas obtenidas son:
Método A: $5, 6, 7, 7, 6, 8, 5, 9, 6, 7$
Método B: $3, 9, 4, 10, 5, 8, 3, 10, 6, 7$

1. Calcula la media y desviación típica de cada grupo.
2. Calcula el CV de cada grupo.
3. ¿Qué método produce resultados más homogéneos?
4. ¿Qué método tiene mayor rendimiento medio?

Problema 3 — Índice de masa corporal

Se recogen el peso (x, en kg) y la altura (y, en m) de 8 personas:

Peso x	55	60	65	70	75	80	85	90
Altura y	1,60	1,63	1,68	1,72	1,75	1,78	1,80	1,82

1. Representa la nube de puntos.
2. Calcula el coeficiente de correlación e interprétalo.

3. Halla la recta de regresión $y = a + bx$.

4. Estima la altura de una persona que pesa 73 kg.

12. Soluciones de las actividades seleccionadas

Solución

Actividad 4.1 — apartado 3 (salarios)
Media: $\bar{x} = (1200+1350+1400+1450+1500+1550+8000)/7 = 16450/7 \approx 2350$€
Mediana: dato central (posición 4 de 7 ordenados) = 1450€.
La **mediana** describe mejor el salario típico porque la media está distorsionada por el valor atípico de 8000€ (seguramente un directivo). La mayoría cobra en torno a 1450€.

Solución

Actividad 6.1 — apartado 1
Datos: $2, 4, 4, 4, 5, 5, 7, 9$. $n = 8$.
$\bar{x} = (2+4+4+4+5+5+7+9)/8 = 40/8 = 5$
$\overline{x^2} = (4+16+16+16+25+25+49+81)/8 = 232/8 = 29$
$s^2 = 29 - 25 = 4; \quad s = 2$

Solución

Actividad 8.1 — apartado 3
Ana: $z_1 = (85-80)/8 = 0{,}625$; $z_2 = (72-68)/10 = 0{,}4$; $\bar{z}_{\text{Ana}} = 0{,}5125$
Luis: $z_1 = (90-80)/8 = 1{,}25$; $z_2 = (68-68)/10 = 0$; $\bar{z}_{\text{Luis}} = 0{,}625$
Luis tiene mejor puntuación tipificada media.

13. Resumen del tema

Lo esencial del Tema 6

- **Variables**: cualitativas (nominal/ordinal) o cuantitativas (discreta/continua).
- **Frecuencias**: absoluta f_i, relativa $h_i = f_i/n$, y sus acumuladas F_i, H_i.
- **Gráficos**: barras (discreta), histograma (continua), ojiva (acumulada), sectores (cualitativa).
- **Posición central**: media $\bar{x} = \sum x_i f_i/n$, mediana ($50,\%$ acumulado), moda (máxima frecuencia).
- **Cuantiles**: Q_1, Q_2, Q_3; RIC $= Q_3 - Q_1$. Diagrama de caja y bigotes.
- **Dispersión**: rango, varianza s^2, desviación típica s, coeficiente de variación $\mathrm{CV} = s/\bar{x} \times 100\,\%$.
- **Forma**: asimetría (γ_1) y curtosis (γ_2).
- **Tipificación**: $z_i = (x_i - \bar{x})/s$ para comparar entre distribuciones y detectar atípicos.
- **Correlación**: $r \in [-1, 1]$; correlación $\neq$ causalidad.
- **Regresión lineal**: $y = a + bx$ con $b = \mathrm{Cov}(X,Y)/s_X^2$, $a = \bar{y} - b\bar{x}$.

¡Enhorabuena por completar el Tema 6!